Classifying author intention for writer feedback in related work

Casey, A.

INCOMA
2019

þÿ C a s e y , A , W e b b e r , B & G B o w a c k a , D 2 0 1 9 , C l a s s i f y i n g a u t h o r i n t e r
feedback in related work . in G Angelova , R Mitkov , I Nikolova & I Temnikova (eds) ,
Proceedings of Recent Advances in Natural Language Processing . INCOMA , Shoumen ,
pp. 178-187 , Recent Advances in Natural Language Processing , Varna , Bulgaria ,
02/09/2019 . https://doi.org/10.26615/978-954-452-056-4_021

http://hdl.handle.net/10138/313131
https://doi.org/10.26615/978-954-452-056-4_021

publishedVersion

Classifying Author Intention for Writer Feedback in Related Work

Arlene Casey
School of Informatics
University of Edinburgh
Edinburgh, UK
a.j.casey@sms.ed.ac.uk

Bonnie Webber
School of Informatics
University of Edinburgh
Edinburgh, UK
bonnie@inf.ed.ac.uk

Dorota Głowacka
Dept. of Computer Science
University of Helsinki
Helsinki, Finland
glowacka@cs.helsinki.fi

Abstract

The ability to produce high-quality publishable material is critical to academic success but many Post-Graduate students struggle to learn to do so. While recent years have seen an increase in tools designed to provide feedback on aspects of writing, one aspect that has so far been neglected is the *Related Work* section of academic research papers. To address this, we have trained a supervised classifier on a corpus of 94 *Related Work* sections and evaluated it against a manually annotated gold standard. The classifier uses novel features pertaining to citation types and co-reference, along with patterns found from studying *Related Works*. We show that these novel features contribute to classifier performance with performance being favourable compared to other similar works that classify author intentions and consider feedback for academic writing.

1 Introduction

Argument structures are key in allowing an author to construct a persuasive message that realizes the author's intention. The automatic identification of such intentions has been shown to be a valuable resource in areas such as summarising information (Teufel and Moens, 2002; Cohan and Goharian, 2015), and understanding citation function and sentiment (Teufel et al., 2006; Jurgens et al., 2018). Recent years have seen more academic writing tools focused on content that use an understanding of expected author intentions to assist in feedback. This is an important resource for Post-Graduate (PG) students who struggle to gain the necessary skills in academic writing that are critical to their success (Aitchison et al., 2012; Paltridge and Starfield, 2007). Automating understanding of author intentions is challenging as research articles, whilst classed in their own genre, are known to differ in content and linguistic style across disciplines (Hyland, 2008).

Despite these challenges, previous work using author intentions has been successful in automating writing feedback though largely focused on the Abstract (Feltrim et al., 2006) or the Introduction (Cotos and Pendar, 2016; Abel, 2018). One reason for this focus on a single section of a research paper is that each section has its own purpose, which encourages different linguistic practices. Existing tools may concentrate on the Introduction due to formative work done by Swales (1990). Swales was one of the first to recognise these intentions in academic writing calling them *rhetoric intentions*. Swales showed how linguistic patterns in the Introduction could be matched to intentions, such as *establishing a territory* or *defining the problem*.

One section of academic papers that has not yet been explored for automated content feedback is the *Related Work* section. One particular challenge students have when learning to write is to find and project their own viewpoint (Kamler and Thomson, 2006). Often their lack of experience and confidence in projecting their own voice amongst established scholars results in students making bland statements about others' work. Such bland statements in the *Related Work* section do nothing more than provide a list of work with no real critical commentary or attempt to relate it to the author's own work. We address this gap of content feedback for *Related Work* by building a model of author intentions that one expects to find in *Related Work*.

We show how the labels of an annotated corpus for author intention, designed for writer feedback in *Related Work*, can be identified reliably. We build on existing methods for feature representa-

Proceedings of Recent Advances in Natural Language Processing, pages 178–187,
Varna, Bulgaria, Sep 2–4, 2019.
https://doi.org/10.26615/978-954-452-056-4_021

tion of author intentions and show that the novel features we introduce contribute significantly to the classifier performance, improving on performance of existing writer feedback tools.

2 Related Work

Automating Author Intentions – Previous models of author intentions in research articles have been successfully automated. One of the first and widely used is Teufel (1999) who proposed Argument Zoning (AZ) which labels sentences with zones representing the rhetoric purpose (author intent) within the global context of a document e.g. background, aim or conclusion. Further work has applied this schema to biology papers (Mizuta and Collier, 2004), with a modified, finer grained approach applied to papers on chemistry (Teufel et al., 2009). Liakata et al. (2012) took a different approach to labelling author intentions, studying the conceptual structure of biology articles treating the article as an investigation. Fisas et al. (2015) develop a schema based on both Liakata and Teufel's work to represent scientific concepts that appear in computer graphics articles. These works have successfully identified author intentions, but they differ from our work by seeking intentions in a global context across a whole article. For example, AZ was developed to support summarisation and information access. The author intentions that these activities would be associated with are rarely found in a *Related Work* section and are unlikely to be helpful in writing feedback for this section. They are nevertheless useful in supporting writing feedback on Abstracts and summaries of PhDs (Feltrim et al., 2006).

Related Work does have in common with other sections the fact that it should contain citations. Understanding the motivations or function of a citation can help determine an author intention (Teufel et al., 2006). Work on citation function has been an area of research for several decades (Weinstock, 1971; Oppenheim and Renn, 1978; Teufel et al., 2006; Angrosh et al., 2012), with more recent work considering how this recognition can be automated. Jurgens et al. (2018) investigates the framing of citations and how this can be used to study the evolution of a field. Teufel et al. (2006) work on automated recognition of citation function and show a strong relationship between function and sentiment. One work that specifically looks at context identification of sentences in *Related Work* is (Angrosh et al., 2010). This work focuses on sentences in terms of their ability to support intelligent information retrieval in digital library services. While aspects of this work and the previous works on citation function are relevant, what is missing is the need to identify where an author talks about their own work in context to other work, showing why it is different or how it fills a gap. As discussed in the Introduction, one of the problems with poor writing in *Related Work* is bland statements that provide lists of citations. Cited works should be ones that have implications for the author's work (Maxwell, 2006). To provide such feedback, we must capture this context in addition to citation function.

Recognising Author Intentions – Specific phrasing has been shown to function in structuring discourse by guiding readers through a text (Hyland, 2012) and can be found to align to sections, such as the Introduction or Results. Most previous work in automating author intentions have utilised these patterns as part of their feature set. The early work of Teufel (1999) (in the domain of computational linguistics) uses cue phrases and lexical patterns that involve parts of speech and citation markers as features. Jurgens et al. (2018) shows how applying bootstrapping to Tuefel's lexicon improves citation function recognition.

Verbs have been shown to have a role in understanding citation function by determining rhetorical and semantic levels. Verbs used to report can show positive and negative aspects of evaluation in cited works and differentiate between intentions e.g in Angrosh et al. (2010) they use reporting verbs that describe (examine, propose), refer to an outcome (develop, show) or show a strength (improve). Citation forms (Swales, 1990) of integral and non-integral have been shown to be a contributing feature to author intention recognition, with studies of novice writers showing that they use a limited range of citation types (Thompson and Tribble, 2001). Our approach also uses linguistic patterns, verb types and citation types to support building our feature set, and we do this within one domain, computational linguistics (cf. Section 4).

Automated Assessment of Writing – Existing, academic writing tools have focused on identifying author intentions, such as those described by Swales (1990), that can be found in an Introduction (Cotos and Pendar, 2016; Anthony and

V. Lashkia, 2003; Abel, 2018). The Criterion online writing service, focuses on automated persuasive essay evaluation and uses recognition of discourse elements based on aspects such as supporting ideas, introductions and conclusion (Burstein et al., 2003, 2004). Several other works have focused on identifying argument components and relations and how these relate to essay scores (Ghosh et al., 2016; Song et al., 2014). Recognizing argument components in this case focuses on premises and claims largely based on the Toulmin model of argumentation (Toulmin, 2003) which is a different approach to ours. In addition, all this work focuses on feedback for persuasive essays which will differ in linguistic practices found in scientific papers and from the author intention structure of a *Related Work*. Overall, whilst aspects may be relevant in general, these methods would not facilitate the kind of content feedback that would help a writer with *Related Work*.

3 Author Intentions to Support Feedback

3.1 Annotation Schema for Data

The need for annotated data is something that previous methods have in common, each using an annotation schema that supports the intentions they seek. It is known that annotation schemas benefit from being task-orientated (Guo et al., 2010). We use an annotation schema developed to recognise author intentions in *Related Work* sections and provide authors with useful feedback (Casey et al., 2019). This schema uses qualities that should be present in *Related Work* sections, following (Kamler and Thomson, 2006). Qualities group into four areas: **Background** – helps the author locate their work in the field, demonstrating they understand their field and its history through indicating seminal works and relevant research fields; **Cited works** – in addition to generally identifying the field, the author should demonstrate specifically (i) which works are most pertinent to their work, (ii) how these works have influenced them and (iii) if and how the current works build on or use these methods; **Gap** – make clear what the gap is and what has specifically not been addressed; **Contribution** – having exposed the gap, the author should identify their contribution. This schema tries to isolate neutral citations that provide mere description, compared to those that highlight gaps or problems, along with identifying where an author talks about their own work and how this relates to the cited work or background in general. The sentence label schema we use can be found in Table 1, and we indicate which of the qualities each label falls into.

3.2 Annotated Dataset

The annotated dataset in (Casey et al., 2019) is composed of papers from the ACL anthology (Bird et al., 2008) that have been pre-annotated for citations and co-reference to the author's own work by (Schäfer et al., 2012). We use 94 papers with *Related Work* sections after removing one due to OCR issues. All papers were conference papers 6-8 pages in length. The authors report annotator agreement, based on Cohen Kappa (Cohen, 1960) at 0.77, which increased to 0.85 following a round of discussion.

3.3 Challenges and Changes to the Schema to Support Automation

Previous works have largely been based on annotating at a sentence level but some works have considered a smaller discourse unit such as (Shatkay et al., 2008). This smaller discourse unit does allow for instances where an author may encode two intentions in one sentence. The data we use is labelled on a sentence basis. The authors in (Casey et al., 2019) acknowledge that this may not be satisfactory as some sentences could be multi-labelled, such as where an author highlights a gap then state their contribution. Just as this is challenging for an annotator to label, it may be more so for an automated classifier.

From the annotated data, some categories were rare and were collapsed into more frequent categories. The distinction these rare labels offered was not necessary. In particular, the two labels that denoted the author said something positive about a citation/field were merged into the appropriate cited/field evaluation categories. We merged the two categories (author's work builds on/adapts/uses X, and author's work is similar to X) into one category. Finally, comparison of two cited works was merged to cited work description. Table 2 shows the distribution of the labels in the 94 *Related Work* sections. We abbreviate some of the labels from the original schema.

Quality	Label	Description
Background	BG-NE	Description of the state of the field, describing/listing known methods or common knowledge. No evidence i.e. citation is not included
	BG-EP	As above but evidence provided i.e.citation included
	BG-EV	Positive, shortcoming, problem or gap in the field
Cited Works	CW-D	Describes cited work, this could be specific details, or very high level details or nothing more than a reference for further information
	A-CW-U	Author's work uses/builds/similar to a cited work
	CW-EV	Positive, shortcoming, problem or gap about the cited work
Gap and Contribution	A-D	Author describes their work with no linguistic marking to other's work or being different
	A-GAP	Author specifically says they address a gap or highlights the novelty of their work
	A-CW-D	Author's highlights how their work is different to cited work
	TXT	Sentence provides information about what will be discussed in the next section

Table 1: Sentence Labels

Sentence Label	Count
BG-EV	90
BG-NE	257
BG-EP	171
CW-EV	133
CW-D	707
A-CW-U	59
A-D	107
TXT	21
A-CW-D	151
A-GAP	59
Total	1755

Table 2: Label class distribution

4 Methods

4.1 Classifiers

All models are trained using LibSVM (Chang and Lin, 2011) with a linear kernel and default settings. SVM's are known to be robust to over-fitting, and perform well in document classification tasks when features are sparse and the set of them is large. SVM does not assume statistical independence, making it a more suitable method when features may be overlapping or interdependent. Initially, we experimented with decision trees. However, when we tested multiple iterations for reliability, both Random Forest (Breiman, 2001) and C4.5 (Sumner et al., 2005) were not only consistently lower in performance (12%) but rare categories showed large variation (15%) between iterations and in some instances labels would not classify. We believe this was due to feature overlap and the problem previously discussed with some labels being multi-class. Due to the unreliability of its performance, we did not pursue the decision tree approach further.

4.2 Features

Features were motivated by other works that classify author intention and citation function, and were extracted on a sentence basis. We use a vector of sentence features as the input to our classifier. The following list summarises the features used in this work.

- **Structural** Positional information, such as relative sentence position, has been useful for identifying background sentences, as these are more likely to occur in an Introduction or *Related Work* than a Results section (Teufel, 1999; Jurgens et al., 2018; Liakata et al., 2012). We do not include relative sentence position but instead use a binary indicator for paragraph start and end sentence, manually added from the original PDF. This is similar to the feature in (Teufel and Moens, 2002) of paragraph structure. We expect this to work like a sentence relative position, as many background statements will come at the start of paragraphs, and towards the end of paragraphs authors will be more likely to relate their own work.

- **Citations Type Features** Authorial and parenthetical citations (Swales, 1990) have been shown to be useful in determining author intention. We build a parser to identify three types of citation: (i) those that form part of the syntax of the sentence (authorial); (ii) those that refer to the name of a system or known algorithm; and (iii) those that provide supporting evidence, found in parenthetical with no syntax e.g **in (Smith, 1990)** although in parenthesis would be of type (i). This slightly differing approach we believe

will help to discriminate between background sentences with citation evidence and citation description sentences. We also take a count of type 1 and 2 citations and a separate count of type 3 citations.

- **N-grams** Work based on a much larger corpus than ours show that n-grams contribute significantly to the performance of their classifier. Liakata et al. (2012) show a 40% contribution and Cotos and Pendar (2016) work is mainly based on n-gram features of 650 Introductions. While our corpus is much smaller (Related Work from 94 articles), we nevertheless include binary values for bigrams and trigrams occurring with a frequency of $\geq$5. We do not remove stop words.

- **Co-referencing Features** Often discussion about a work or the author's work will be carried out over several sentences. The later sentences can have co-references to the original citation such as *'this paper' 'this model'* However, as Teufel (1999) shows, determining what she calls agents (e.g US_AGENT - 'our paper'), these co-reference phrases can be ambiguous. For example does *'this paper'* mean the previously cited paper or it is referencing the author's work. We take a different approach and use the annotations in our corpus for (i) references to the authors own work, (ii) cited work. In addition, we manually mark co-reference to multiple works in background sentences e.g *'previous work has been done in the area of ..'* and co-reference to previously cited work e.g. *'these previously mentioned works above'*

- **Verb Features** We use part of speech (POS) tags to identify verbs, treating the six possible VB tags (VB, VBD, VBG, VPN, VBP, VBZ) as binary features of being present or not in a sentence. Having substituted the co-references, described above, in a sentence we then parse for dependency extracting subject and object verb pairs in every sentence.

- **Linguistic Patterns** Teufel (1999) makes available a list of patterns containing cue phrases/words, lexical categories, constrained by PoS tags, developed on computational linguistic literature. Like (Jurgens et al., 2018) we use this list and adapt it manually using patterns we observe in *Related Work*. For example, one aspect we consider is contrasts that occur at the beginning of a sentence and those that happen mid sentence, creating lexical expressions to capture these. We also produce finer grained lexicon patterns for discourse connectives as these are indicative of a continuation sentence. Within those patterns we include citation types and co-references as described above.

- **Sentiment** We use our adapted version of Teufel's list to identify positive and negative words (e.g *advantageous* - positive adjective, *inaccurate* - negative adjective). In addition, we use the 82 polar phrases found in (Athar, 2014). We parse each sentence and count the positive and negative words.

- **Counts** Counts of sentence words, nouns, adverbs, discourse connectives.

- **Subject** We assign a sentence subject label before assigning a sentence label to decide if the sentence is about a citation, background or field information, author's work, or a combination of author's work and cited work. This subject feature is based on rules of sentence and previous sentence features e.g our finer grained approach to discourse connectives in conjunction with co-reference markers help us to understand subject.

5 Experimental Setup and Evaluation

5.1 Baseline

We provide two baselines, one with n-gram features only and one with all features based on the majority class.

5.2 Evaluation

Our work is similar to other automated classifications but not directly comparable as schemas and experimental settings differ. Our results are more comparable to the works of (Teufel, 1999; Jurgens et al., 2018; Teufel and Kan, 2011) as we use the same pattern list from (Teufel, 1999) as a starting point. These works use Naive Bayes, Random Forest and Maximum Entropy as classifier methods. We report their published Macro F1 scores, range of F1 scores for labels and the number of labels in the schema for comparison (Table

Features	Prec	Recall	F1	Acc%
ALL	.69 (.005)	.7 (.004)	.7 (.005)	70 (.48)
Cotos(2016)	.686	.549	.61	72.9%
Teufel(2011)	.478	.376	.4142	66.8%

Table 3: Classifier Performance, Mean scores after 10 iterations, Variance in brackets

System	**Macro F1/Range**	**Label No.**
(Teufel and Kan, 2011)	0.41 (0.19-0.81)	8
(Jurgens et al., 2018)	0.53	6
(Teufel, 1999)	0.68 (0.28-0.86)	12
(Cotos and Pendar, 2016)	0.61 (0.36-0.85)	17
Our Work		
-all feat	0.70 (.25 -0.88)	10
- no novel feat	0.54 (.15-.87)	
Baseline		
Ngram(B,T)	0.39 (.02-.68)	
Majority	0.57 (-)	

Table 4: Classifier Comparison, * significant 0.01

4). Also included is the work of (Cotos and Pendar, 2016) which focuses on writer feedback for Introductions. This is a much larger corpus using 650 annotated Introductions but fewer features, focusing on unigram and trigrams. However, it also uses SVM for classification. Where available, we also report Precision, Recall and Accuracy from these works to compare against our best performing model in Table 3.

Reliability of our model is important to ensure consistent results. Therefore, in addition to 10-fold cross validation, we carry out 10 iterations of the All features model, reporting on mean precision, recall, F1, Accuracy and variance in Table 3. None of our iterations produced significantly different results, demonstrating reliability and low variation. Significance, where noted, is tested with corrected t-test, p $<$0.01, (Nadeau and Bengio, 1999).

We also look at the features in our model and how they influence the label F1 scores with leave one out (LOO), which highlights the performance decrease when a single feature is omitted and single features (SF), which highlights the contribution of a single feature to performance. Looking at individual label features is important as having just one label perform poorly, such as being able to recognise an author gap sentence or where an author says how their work is different, will impact our goal of giving writer feedback.

6 Results

6.1 Classifier Performance

We compare our results to those mentioned in Section 5, Table 4. Comparing F1 scores overall, we outperform the other systems by a reasonable margin. In addition, the range of F1 scores for our labels are also similar to other systems. We outperform the work of Cotos and Pendar (2016), who looks at classification for Introduction feedback, despite their work being based on a bigger annotated corpus. We significantly outperform both our baselines of n-grams and majority class. We rerun our classification (no novel feat) removing the manual additions we added to the original pattern list of Teufel (1999), removing co-references and subject labels. This results in lower performance, significant (p $<$0.01) than our all features and our majority baseline.

6.2 Feature Contribution

Here we examine feature contributions by single feature and leave one out cross validation runs (Table 5). For each category, we highlight the lowest score in bold, which corresponds to the feature being left out. We place in brackets any scores higher than the All features model.

More frequently occurring categories, cited work description (CW-D), background sentences with and without evidence (BG-NE, BG-EP) are more robust to feature omissions. Features are not independent, so many of the patterns cover the n-gram features which may be why leaving out n-grams has less impact than expected. In the lower part of the table, n-grams as a single feature contributes most to labels TEXT and CW-DESC. Compared to other works that have used n-grams, our size is much smaller at $<$3000, whereas Liakata et al. (2012) used $\sim$42000 and Cotos and Pendar (2016) had $\sim$27000. It would be expected in a much larger corpus that n-grams will contribute more as a feature.

The removal of the paragraph start and end markers makes relatively little difference, with the exception of the author gap category (A-GAP): Being a rare category, this addition although small is important. Sentiment contributes in a small way to performance but particularly in the evaluation labels (BG-EV, CW-EV) as expected. Surprisingly, sentiment contributes to the text label. However, within text-labelled sentences, both these counts are zero, which may explain why it con-

Features	BG-EV	BG-NE	BG-EP	CW-EV	CW-D	A-CW-U	A-D	TXT	A-CW-D	A-GAP
ALL	.39	.72	.73	.53	.84	.48	.47	.88	.63	.25
Feat-(LOO)										
-subject	**.33**	.62	.71	.51	**.81**	.49	.41	.85	.64	**.22**
-n-grams	.33	**.70**	**.70**	.53	.84	(.50)	**.39**	.83	.62	.25
-verbtense	.35	.71	.72	.51	.84	.48	.46	.88	**.66**	.32
-sentiment	.34	.71	.71	**.50**	.84	**.46**	.43	**.67**	**.61**	(.28)
-counts	(.40)	.72	.73	.52	.84	(.50)	.46	.87	(.64)	.26
-Tot cit count	.38	.71	(.74)	(.54)	(.85)	.49	(.48)	.88	.64	.26
-paragraph	(.40)	.71	.73	(.54)	.84	.49	.47	.87	.62	**.22**

Features	BG-EV	BG-NE	BG-EP	CW-EV	CW-D	A-CW-U	A-D	TXT	A-DIFF	A-GAP
ALL	.39	.72	.73	.53	.84	.48	.47	.88	.63	.25
Feat-(SF)										
-patterns	.30	.54	(.74)	.41	.77	(.57)	(.48)	.80	(.65)	.26
-subject	-	.58	-	-	.80	-	.45	.75	.46	-
-sub+patt+dep	.31	.72	.73	.47	.83	(.55)	.46	.84	.63	(.27)
-n-grams	.11	.31	.21	.24	.62	.23	.04	.68	.39	.02

Table 5: F-Measures for Features and Labels,10 cross validation

tributes here. Neither of our evaluations labels (BG-EV, CW-EV) perform as well as expected. These two labels are merged from the annotation schema, positive and shortcoming/problem into one evaluation label. These original labels are both quite different linguistically and we speculate that this might prove difficult for the classifier.

Total citation counts and counts of adverbs, words, nouns and discourse connectives seem to actually make the performance of the classifier worse on many of the labels, although not significantly so. There is an overlap in total citation counts with the count of our citation types perhaps indicating this feature could be omitted.

We note that the features we add to the pattern list, dependencies and subject label show very close to performance of the All features model. We observe better performance on the rare label author gap (A-GAP) with just these features alone.

Most categories are negatively impacted by the removal of the subject label with the exception of author uses/build/similar to cited work (A-CW-U) and authors work differs from cited work (A-CW-D). As a single feature we see that subject is important to the classifier performance and contributes to several of the labels - background with no evidence (BG-NE), cited work description (CW-D), author description (A-D) and author and cited work differ (A-CW-D). Leaving out subject label was the only feature to cause a drop in classifier performance that was significant. In Table 6 and Table 7 experiments from using a gold subject label and using a history feature of previous label are presented. History label was previously shown by (Liakata et al., 2012) to contribute to sentence classification. Our gold subject label was determined from the annotated label. We see that determining this label accurately has an almost 15% increase in the performance of our classifier and an increase in F1 score for all label categories. Including a previous label also increases the classifier performance, but this increase was not significant.

7 Discussion and Conclusions

We use a manually annotated data-set designed for support of writer feedback of a *Related Work* section and show that we can outperform existing similar methods. We describe our feature set, proposing some novel features such as co-reference specific to *Related Works*, citation types and include these in our adapted pattern set. We show the introduction of our features over and above the original pattern features (Teufel, 1999) was a contributing factor to the performance of our classifier. This highlights the importance of understanding the author intentions of interest and looking for patterns that are specific to these. This major contribution of patterns though is also a limitation in that this is built on a study of patterns that occur within the computational linguistic domain and how it would perform in another domain remains to be investigated. Recent work of (Asadi et al., 2019) show that using WordNet roots for Nouns, e.g where nouns are taken to their more general form (e.g., *mm* and *cm* become *quantity*, is a useful feature for author intention identification. The application of WordNet is one possible

Features	BG-EV	BG-NE	BG-EP	CW-EV	CW-D	A-CW-U	A-D	TXT	A-D	A-G
ALL	.39	.72	.73	.53	.84	.48	.47	.88	.63	.25
+Plabel	.50	.60	.70	.60	.86	.51	.46	.63	.61	.27
+GoldSubject	.61	.86	.88	.67	.94	.68	.72	1	.88	.4

Table 6: Mean F-Measures for Labels All features and All with Gold Subject and Previous label

Features	Prec	Recall	F1	Acc%
ALL	.69	.7	.7	70
+Plabel	.71	.72	.71	71.72
+GoldSubject	.84	.85	.84	84.6

Table 7: Classifier Performance, Mean scores after 10 iterations

avenue that may assist in transitioning our pattern list to another domain.

In future work, we intend to investigate augmenting our pattern set further. Jurgens et al. (2018) implement a bootstrapping pattern that identifies over four times the manually curated patterns, identifying new patterns that apply in a citing sentence, the preceding or following sentence. Bootstrapping to expand seed cue phrases based on rhetorical relations (Abdalla and Teufel, 2006) has also been successful. Incorporating more information from a preceding or following sentence we believe could help classify sentences where there is no linguistic clue as to the subject e.g. those that carry on describing a cited work but their is no co-reference to signal the subject. Understanding sentence subject is important, currently it contributes to the classifier performance but we show an almost 15% increase in performance that could occur using a gold sentence subject label. Having a way to improve our current implementation of sentence subject assignment would be beneficial.

Our overall intention for this work is to support writer feedback and so we intend to investigate how well our current level of automatic recognition of author intentions can support feedback and how useful this is to novice writers.

References

Rashid M. Abdalla and Simone Teufel. 2006. A bootstrapping approach to unsupervised detection of cue phrase variants. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Stroudsburg, PA, USA, ACL-44, pages 921–928. https://doi.org/10.3115/1220175.1220291.

Sophie Kitto Kirsty Knight Simon Buckingham Shum Simon Abel. 2018. Designing personalised, automated feedback to develop students research writing skills. In *Proceedings of 2018 Australasian Society for Computers in Learning in Tertiary Education*. pages 15–24.

Claire Aitchison, Janice Catterall, Pauline Ross, and Shelley Burgin. 2012. 'Tough love and tears': learning doctoral writing in the sciences. *Higher Education Research & Development* 31(4):435–447. https://doi.org/10.1080/07294360.2011.559195.

Mandya A. Angrosh, Stephen Cranefield, and Nigel Stanger. 2010. Context identification of sentences in related work sections using a conditional random field: Towards intelligent digital libraries. In *Proceedings of the 10th Annual Joint Conference on Digital Libraries*. ACM, New York, NY, USA, JCDL '10, pages 293–302. https://doi.org/10.1145/1816123.1816168.

Mandya A. Angrosh, Stephen Cranefield, and Nigel Stanger. 2012. A citation centric annotation scheme for scientific articles. In *Proceedings of the Australasian Language Technology Association Workshop 2012*. Dunedin, New Zealand, pages 5–14. https://www.aclweb.org/anthology/U12-1003.

Laurence Anthony and George V. Lashkia. 2003. Mover: A Machine Learning Tool to Assist in the Reading and Writing of Technical Papers. *Professional Communication, IEEE Transactions on* 46:185 – 193. https://doi.org/10.1109/TPC.2003.816789.

Nasrin Asadi, Kambiz Badie, and Maryam Tayefeh Mahmoudi. 2019. Automatic zone identification in scientific papers via fusion techniques. *Scientometrics* 119(2):845–862. https://doi.org/10.1007/s11192-019-03060-9.

Awais Athar. 2014. Sentiment analysis of scientific citations. Technical Report UCAM-CL-TR-856, University of Cambridge, Computer Laboratory. https://www.cl.cam.ac.uk/techreports/UCAM-CL-TR-856.pdf.

Steven Bird, Robert Dale, Bonnie Dorr, Bryan Gibson, Mark Joseph, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir Radev, and Yee Fan Tan. 2008. The ACL Anthology Reference Corpus: A Reference Dataset for Bibliographic Research in Computational Linguistics. In *LREC 2008*.

Leo Breiman. 2001. Random forests. *Machine learning* 45(1):5–32.

J Burstein, D Marcu, and K Knight. 2003. Finding the WRITE stuff: Automatic identification of discourse structure in student essays. *IEEE Intelligent Systems* .

Jill Burstein, Martin Chodorow, and Claudia Leacock. 2004. Automated Essay Evaluation: The Criterion Online Writing Service. *AI Magazine* 25:27–36.

Arlene J Casey, Bonnie Webber, and Dorota Głowacka. 2019. A Framework for Annotating 'Related Works', to Support Feedback to Novice Writers. In *LAW '13: Proceedings of the Linguistic Annotation Workshop*. Association for Computational Linguistics, Stroudsburg, PA, USA.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* 2:27:1–27:27.

Arman Cohan and Nazli Goharian. 2015. Scientific article summarization using citation-context and article's discourse structure. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, pages 390–400. https://doi.org/10.18653/v1/D15-1045.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20(1):37–46.

Elena Cotos and Nick Pendar. 2016. Discourse classification into rhetorical functions for awe feedback. *calico journal* 33(1):92–116.

Valéria D Feltrim, Simone Teufel, Maria Graças V das Nunes, and Sandra M Aluísio. 2006. Argumentative zoning applied to critiquing novices scientific abstracts. In *Computing Attitude and Affect in Text: Theory and Applications*, Springer, pages 233–246.

Beatriz Fisas, Horacio Saggion, and Francesco Ronzano. 2015. On the discoursive structure of computer graphics research papers. In *Proceedings of the 9th linguistic annotation workshop*. Association for Computational Linguistics, Denver, Colorado, USA, pages 42–51. https://doi.org/10.3115/v1/W15-1605.

Debanjan Ghosh, Aquila Khanam, Yubo Han, and Smaranda Muresan. 2016. Coarse-grained argumentation features for scoring persuasive essays. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Berlin, Germany, pages 549–554. https://doi.org/10.18653/v1/P16-2089.

Yufan Guo, Anna Korhonen, Maria Liakata, Ilona Silins Karolinska, Lin Sun, and Ulla Stenius. 2010. Identifying the information structure of scientific abstracts: an investigation of three different schemes. In *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*. Association for Computational Linguistics, pages 99–107.

Ken Hyland. 2008. As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes* 27(1):4–21.

Ken Hyland. 2012. Bundles in academic discourse. *Annual Review of Applied Linguistics* 32:150–169. https://doi.org/10.1017/S0267190512000037.

David Jurgens, Srijan Kumar, Raine Hoover, Dan McFarland, and Dan Jurafsky. 2018. Measuring the Evolution of a Scientific Field through Citation Frames. *Transactions of the Association of Computational Linguistics* 6:391–406.

Barbara Kamler and Pat Thomson. 2006. *Helping doctoral students write: Pedagogies for supervision*. Routledge. https://doi.org/10.4324/9780203969816.

Maria Liakata, Shyamasree Saha, Simon Dobnik, Colin Batchelor, and Dietrich Rebholz-Schuhmann. 2012. Automatic recognition of conceptualization zones in scientific articles and two life science applications. *Bioinformatics* 28(7):991–1000.

Joseph A. Maxwell. 2006. Literature Reviews of, and for, Educational Research: A Commentary on Boote and Beile's "Scholars Before Researchers". *Educational Researcher* 35(9):28–31. https://doi.org/10.3102/0013189X035009028.

Yoko Mizuta and Nigel Collier. 2004. Zone identification in biology articles as a basis for information extraction. In *Proceedings of the international joint workshop on natural language processing in biomedicine and its applications*. Association for Computational Linguistics, pages 29–35. https://www.aclweb.org/anthology/W04-1205.

Claude Nadeau and Yoshua Bengio. 1999. Inference for the Generalization Error. In *Proceedings of the 12th International Conference on Neural Information Processing Systems*. MIT Press, Cambridge, MA, USA, NIPS'99, pages 307–313. http://dl.acm.org/citation.cfm?id=3009657.3009701.

Charles Oppenheim and Susan P Renn. 1978. Highly cited old papers and the reasons why they continue to be cited. *Journal of the American Society for Information Science* 29(5):225–231.

Brian Paltridge and Sue Starfield. 2007. *Thesis and Dissertation Writing in a Second Language*. Routledge. https://doi.org/10.4324/9780203960813.

Ulrich Schäfer, Christian Spurk, and Jörg Steffen. 2012. A fully coreference-annotated corpus of scholarly papers from the ACL anthology. In *Proceedings of COLING 2012: Posters*. The COLING 2012 Organizing Committee, Mumbai, India, pages 1059–1070. https://www.aclweb.org/anthology/C12-2103.

Hagit Shatkay, Fengxia Pan, Andrey Rzhetsky, and W John Wilbur. 2008. Multi-dimensional classification of biomedical text: toward automated, practical provision of high-utility text to diverse users. *Bioinformatics* 24(18):2086–2093. https://doi.org/10.1093/bioinformatics/btn381.

Yi Song, Michael Heilman, Beata Beigman Klebanov, and Paul Deane. 2014. Applying argumentation schemes for essay scoring. In *Proceedings of the First Workshop on Argumentation Mining*. Association for Computational Linguistics, Baltimore, Maryland, pages 69–78. https://doi.org/10.3115/v1/W14-2110.

Marc Sumner, Eibe Frank, and Mark A Hall. 2005. Speeding up logistic model tree induction. *PKDD* LNCS 3721:675–683. https://hdl.handle.net/10289/1446.

John M Swales. 1990. *Genre Analysis: English in academic and research settings*. Cambridge University Press.

Simone Teufel. 1999. *Argumentative zoning: Information extraction from scientific text*. Ph.D. thesis, University of Edinburgh.

Simone Teufel and Minyen Kan. 2011. Robust Argumentative Zoning for Sensemaking in Scholarly Documents. In *In Advanced Language Technologies for Digital Libraries*. Springer, pages 154–170.

Simone Teufel and Marc Moens. 2002. Summarizing scientific articles: Experiments with relevance and rhetorical status. *Computational Linguistics* 28(4):409–445. https://doi.org/10.1162/089120102762671936.

Simone Teufel, Advaith Siddharthan, and Colin Batchelor. 2009. Towards domain-independent argumentative zoning: Evidence from chemistry and computational linguistics. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Singapore, pages 1493–1502. https://www.aclweb.org/anthology/D09-1155.

Simone Teufel, Advaith Siddharthan, and Dan Tidhar. 2006. Automatic classification of citation function. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Sydney, Australia, pages 103–110. https://www.aclweb.org/anthology/W06-1613.

Paul Thompson and Chris Tribble. 2001. Looking at citations: Using corpora in english for academic purposes. *Language Learning Technology* 5(3):91 – 105.

Stephen E Toulmin. 2003. *The Uses of Argument*. Cambridge University Press.

Melvin Weinstock. 1971. Citation indexes. encyclopedia of library and information science. volume 5. eds. a. kent & h. lancour.